基于多粒度特征融合模型的多说话人声纹识别研究

李娅[1],周斌[1*],胡波[2]
(1 中南民族大学 计算机学院,武汉 430074;2 武汉东信同邦信息技术有限公司,武汉 430074)

摘 要 声纹识别是一种通过比对语音中的说话人特征来确认身份的技术.其流程包括从不定长的单人语音片段中提取声学特征,再由模型提取说话人特征,最后通过余弦相似度判断是否为同一说话人.多说话人声纹识别则需从包含多位说话人的音频中识别各自身份,该任务的关键在于提取具判别力的说话人特征.为此提出了多粒度特征融合模型(MGFF-TDNN),在MGFF-TDNN中,首先使用二维深度可分离卷积模块(DSM)作为前端特征提取器,以增强时频域特征的建模.其次,为了实现多粒度特征融合,设计了多粒度融合结构(M-TDNN),采用时延神经网络和音素级特征池化来捕获不同粒度的信息.在VoxCeleb数据集上的实验表明:MGFF-TDNN模型在使用了更少的参数量(4.78M)和计算资源(1.49G的浮点计算量)的情况下,仍表现出良好的性能.

关键词 声纹识别;多说话人;多粒度特征融合;深度可分离卷积
中图分类号 TP391.4 **文献标志码** A **文章编号** 1672-4321(2026)01-0077-09
doi:10.20056/j.cnki.ZNMDZK.20250828

Research on multi-speaker voice recognition based on a multi-granularity feature fusion model

LI Ya[1], ZHOU Bin[1*], HU Bo[2]
(1 College of Computer Science, South-Central Minzu University, Wuhan 430074, China;
2 Wuhan Dongxin Tongbang Information Technology Co., Ltd., Wuhan 430074, China)

Abstract Voice recognition is a technique used to determine a speaker's identity by comparing speaker voice features. The process typically involves extracting acoustic features from variable-length speech segment of single speaker. Then, speaker features are extracted from these acoustic features in the model. Finally, cosine similarity is used to determine whether two voice segments originate from the same speaker or different speakers. In the case of multi-speaker voice recognition, the goal is to identify individual speaker from audio segment containing multiple speakers' s voice, the primary challenge lies in extracting discriminative features of speakers. To address the problem, a Multi-Granularity Feature Fusion model (MGFF-TDNN) is proposed. A two dimensional depth-wise separable convolution module (DSM) is utilized as the front-end feature extractor to enhance modeling of time-frequency domain features. Additionally, to achieve multi-granularity feature fusion, a multi-granularity fusion structure (M-TDNN) is designed, which employs a time-delay neural network and phoneme-level pooling to capture information at various granularities. Experiments on the VoxCeleb dataset demonstrate that the MGFF-TDNN model delivers strong speaker voice recognition performance while utilizing fewer parameters(4.78M) and computational resources(1.49G FLOPs).

Keywords voice recognition; multi-speaker; multi-granularity feature fusion; depth-wise separable convolution

声纹是一种从语音信号中提取出来的表征说话人的特征,声纹识别也可以称为说话人验证(Speaker Verification,SV),是一项旨在验证给定说话人是否来自注册说话人的任务.目前很多声纹识别系统都是针对单个说话人的身份识别,广泛应用于安全身份识别、智能家居和司法鉴定等领域.多

收稿日期 2024-09-18
* **通信作者** 周斌(1971-),男,教授,博士,研究方向:大数据处理,E-mail: binzhou@mail.scuec.edu.cn
基金项目 湖北省技术创新专项基金资助项目(2019ADC071);中央高校基本科研业务费专项资金资助项目(CZY23006)

说话人声纹识别实际上也是基于单说话人的声纹识别,只是在其基础上加上较为复杂的前处理和后处理操作,其核心仍是从语音中提取出具有鲁棒性且判别性强的说话人特征.

随着深度学习技术的飞速发展,声纹识别技术也进入深度学习时代,涌现出许多基于深度学习的模型和方法.d矢量(d-vector)[1-2]是第一个将深度学习应用于SV的矢量,随后是x矢量(x-vector)[3-4],它从可变长度的语音片段中提取固定维度的说话人特征.基于深度学习的SV已逐渐演变为两种主流架构:一种使用时延神经网络(Time-delay Neural Network,TDNN)[5-7]作为主干网络来提取说话人特征,另一种使用残差网络(Residual Network,ResNet)[8-10].

基于TDNN的模型可以有效地提取不同时序的上下文特征,但完全捕获频域特征通常需要增加大量的模型参数.同样,基于ResNet的模型虽然能够同时对时域和频域维度进行建模,但通常难以实现最先进的(State-of-the-Art,SOTA)性能,并且通常需要大量参数和计算资源.例如,一个成功的变体模型ECAPA-TDNN[5]实现了SOTA性能,但是使用了大量滤波器组,带来了高参数量和计算复杂度.ECAPA-TDNN提出SE-Res2Block(Sequeeze and Excitation Res2Block)模块,它使用一维Res2Net[10]模块与挤压激励模块(Squeeze-Excitation Block,SE-Block)[11]相结合的方式来提取帧级特征,实现了令人印象深刻的性能.基于ResNet[12]的模型使用二维卷积来提取时间和频率维度的特征,也表现出较为出色的效果.然而,这些基于ResNet的模型通常需要大量的参数和计算资源才能获得有竞争力的性能.

为解决上述问题,本文提出了MGFF-TDNN(Multi-granularity Feature Fusion TDNN)模型,这是一种基于TDNN的增强架构,更加注重于多粒度特征的上下文建模.首先,为了增强模型在频域的特征建模,采用二维深度可分离卷积模块(Depth-wise Separable Convolution Module,DSM)[13]作为前端提取器来捕获时频域特征;其次,由文献[9,14]启发,设计了一个多粒度时延神经网络(Multi-granularity Temporal Delay Neural Network,M-TDNN)模块.在M-TDNN中,集成了音素级特征池化模块,以增强细粒度特征建模.此外,通过逐渐增加TDNN模块的膨胀因子,以扩展时序上下文的特征建模,然后把这两部分的输出拼接送入到SE模块中,来建模不同粒度特征之间的内部依赖.

1 相关工作

针对多说话人的声纹识别技术,往往使用多阶段的方式分模块处理,因为端到端的方案现阶段存在明显的劣势,例如模型依赖大量的数据集训练,且泛化能力较弱;模型识别人数准确性较差;针对长音频对话数据,模型长序列编解码受限,导致应用场景较弱.相对来说,多阶段处理方案在应用性能上更佳.

1.1 多说话人语音前处理

多说话人语音前处理需要将长对话中不同说话人的语音进行语音片段分离,为后续提取说话人特征做准备.在工程应用上,其一般处理步骤为:对音频数据进行语音活动检测(Voice Activity Detection,VAD),得到语音中的人声片段;然后使用固定长度的时间窗口和偏移对片段进行进一步切分,用于后续的特征提取.这种方式在时间窗口出现两个及以上说话人时识别效果不佳.实际上,多说话人声纹识别往往不是单独作为一个任务,而是结合其他语音任务共同实现一个工程应用.在前端,通过语音识别模型可以得到语音对应的文本数据和每个字对应的时间戳数据,借助语言模型的语义能力对交互语音进行分句,根据分句时间戳得到对应的语音进而进行声纹识别.

本研究在实际应用中发现,语音质量对声纹识别模型的性能有较大影响,多说话人声纹识别作为一个多阶段任务,每一阶段的处理策略都会影响最终呈现的性能,所以需要将语音质量考虑其中.由于语音质量的差异,语音识别的结果也会受到影响,进而影响语言模型的分句能力,无法直接将语言模型的分句结果用于声纹识别,因为分句中仍然存在较大可能性出现多个说话人的表达.因此,需要根据语音质量评估的结果,采用合适的切分策略.

语音质量评估[15]的平均意见得分(Mean Opinion Score,MOS)取值为0~5,一般大于等于3被认为是正常及优良音质,此时可以针对语音识别的文本进行分词处理,再进行说话人特征的提取,MOS值小于3时,则对语音识别的文本进行分句处理用于后续的特征提取,这样的切分策略是为了保证说话人特征提取阶段可以充分提取有效信息且尽量避免用于特征提取的语音包含多个说话人.

1.2 说话人特征提取

1.2.1 提取声学特征

声学特征提取是将音频数据转换为时频域特

征向量，方便后续模型的处理和分析，其流程包括预加重、分帧、加窗、快速傅里叶变换、滤波器、对数、离散余弦变换等．目前声纹识别领域常用的声学特征有梅尔频率倒谱系数（Mel Frequency Cepstral Coefficients，MFCC）和滤波器组（Filter Bank，FBank）特征．两者在前期的计算步骤相同，只是在最后阶段 FBank 特征没有经过离散余弦变换，因此 FBank 包含更抽象的说话人特征[16]．本文使用 FBank 作为模型训练的声学特征．

1.2.2　说话人模型

在当前的声纹识别领域中，主流的框架由帧级特征提取模块、池化模块和语句级特征提取模块组成[3]．在这一领域的不断发展中，ECAPA-TDNN[5]因其卓越的表现而受到广泛关注，逐渐成为研究工作的基准模型．此外，具有类残差结构的二维卷积网络也进入说话人验证领域，Res2Net[8]模型就是一个值得注意的例子．该模型将输入通道划分成几个子分组，每个分组使用单独的卷积核进行特征提取，分组间使用类残差的方式连接起来，以获得不同尺度的特征提取．不同帧数的声学特征经过模型处理后得到固定维度的说话人特征．

1.3　说话人特征后端聚类

经过前几阶段处理，可以从多说话人的音频数据得到大量同维度的说话人特征，表征了某一语音段内的说话人信息．后处理阶段需要将说话人特征向量进行聚类处理，常用的聚类方法有 k 均值聚类和谱聚类[17]．由于谱聚类的算法流程中会对分类数目进行估计，在未知说话人个数的情况下普遍使用谱聚类方法．

由于后端需要进行聚类操作，多说话人声纹识别技术需要保证提取的说话人特征更具有鲁棒性且判别性强，使得不同的说话人特征在高维空间形成良好的聚类效果，同时减少同一说话人的空间差异，从而提高后端聚类操作的准确性和声纹识别的整体性能．

2　MGFF-TDNN 模型

MGFF-TDNN 模型的总体架构如图 1 所示．该模型由两个主要模块组成：深度可分离模块（Depth-wise Separable Convolution Module，DSM）和多粒度时延神经网络（Multi-granularity Temporal Delay Neural Network，M-TDNN）模块．在 DSM 中，通过扩展通道维度，引入了二维深度可分离残差网络层，用于时频域特征的初始建模．随后，生成的特征图沿通道和频率维度展平，并输入到多个 M-TDNNBlocks 中，每个 M-TDNNBlocks 包含不同数量的 M-TDNN 层．每个 M-TDNN 层提取不同粒度的特征，然后通过 SE（Squeeze-Excitation）模块进行融合，以获得不同粒度特征的注意力权重．M-TDNNBlocks 输出说话人的帧级特征，然后通过统计池化层和全连接层得到固定维度的说话人嵌入特征．统计池化层将输入的特征沿着时间维度计算全局的统计特征，得到均值向量和标准差向量，然后将均值和标准差的特征向量拼接起来得到一个固定长度的特征向量，这样得到的特征向量的形状与时间维度无关，因此模型可以处理任意长度的音频．全连接层接收来自统计池化层的固定长度特征向量，为了方便后续计算和处理，将其映射到一个更低维的空间中，一般设置 128、192 或 512 维，本文取 192 维作为最终的说话人嵌入特征维度．

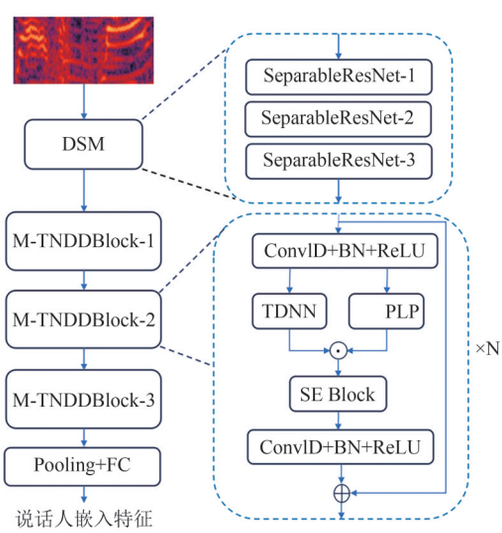


图 1　MGFF-TDNN 模型的总体架构图

Fig. 1　The overall architecture diagram of the MGFF-TDNN model

2.1　DSM 模块

普通的时延神经网络沿着时间维度对声学特征进行卷积操作，捕捉序列中的时域特征，其卷积核虽然完全覆盖了频域特征，但是需要大量的滤波器组才能对频域特征进行充分建模，这无形之中增加了模型的参数量．为了对时域和频域特征进行有效建模，在前端加入二维卷积模块，受文献[13，18]的启发，本文应用了深度可分离模块，该模块使用了多层可分离残差结构，每一层的微结构被称为倒置残差块或纺锤型残差块（Inverted Residual Block）[13]．该结构最初被用在计算机视觉领域，使用深度卷积

(Depth-wise Convolution)和逐点卷积(Point-wise Convolution)组合的方式来提取特征,减少了卷积层的计算复杂度,在基本不影响性能的基础上提高了网络的表达能力,如图2所示,假设输入是x,经过3次卷积操作和残差连接后输出为y,该过程可以用以下公式描述:

$$y = \mathrm{ReLU}\Big(x + \mathrm{BN}\Big(W_3 \cdot \mathrm{ReLU}(\mathrm{BN}\Big(W_2 \cdot \mathrm{ReLU}\Big(\mathrm{BN}\Big(W_1 \cdot x\Big)\Big)\Big)\Big)\Big)\Big), \tag{1}$$

其中W_1和W_3表示point-wise卷积,它们的输出通道分别是$C*t$和C,t是扩展因子,一般取5~10(在本文中设置为6),W_2表示depth-wise卷积,它的输出通道也是$C*t$.BN是批次归一化[19],ReLU(·)是线性整流单元.该模块首先采用逐点卷积来增加通道的维数,然后利用深度卷积进行特征提取,最后采用逐点卷积进行维数还原.这种方法实现了在高维空间进行特征提取,有利于全面捕捉语音信号中的时频域特性,而且,使用二次逐点卷积调整特征维度,确保将模型参数和计算复杂度维持在适当的范围内.

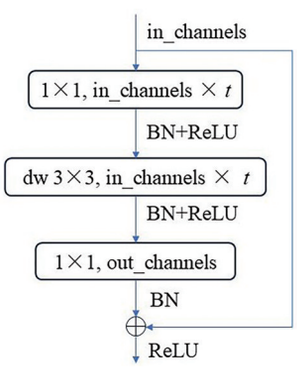


图2 深度可分离模块示意图

Fig. 2 Diagram of depth-wise separable module

2.2 M-TDNN模块

从应用的角度看,说话人验证任务面临的一个较大挑战是对短语音(通常只有3~5s)的说话人建模,与长语音的说话人建模不同,基于短语音的说话人建模要更关注语音的细节信息,即更细粒度的信息.在相同的方法下,从长语音中提取出丰富信息的概率更大,导致人们忽略了细粒度的重要性,而把重心放在是否有足够的上下文上.文献[9]的实验结果实际上也表明了局部特征融合对模型性能的显著贡献.

基于以上的分析和文献[14]的启发,本文提出了多粒度时延神经网络模块——M-TDNN模块(见表1),提取不同粒度的特征信息,然后用一定的融合策略将特征融合起来,获取不同粒度的重要特征.多粒度特征提取与以往所提到的多尺度特征聚合方法[20-21]有所区别,多尺度关注层与层之间的特征融合,是对不同长度的上下文信息进行融合,而多粒度特征融合强调层内部的不同粒度信息之间的融合,避免特征单一的表征方式.

语音特征的时间步是由每个音频帧组成,且帧长一般取25 ms,由于音素单元的时间跨度一般包含多个音频帧,所以本文在M-TDNN层中引入了音素级池化(Phoneme Level Pooling,PLP),它是在滑动窗口上使用标准最大池化来实现的,以捕获细粒度的特征.为了减少信息损失,PLP滑动窗口有50%上下文的重叠,本文实验中滑动窗口的大小设置为8.为了确保PLP的输出张量与输入张量在时间维度上对齐,实验中以跨时间轴的方式复制窗口的池化特征,这有助于后续的特征融合.控制TDNN和PLP的输出维度,并沿时间轴将它们的输出拼接起来,然后通过SE模块对其进行处理.这种方法允许集成具有不同权重的特征,从而增强时频域特征的建模.

如图1所示,假设标记M-TDNN Block中某一层的输入为Y,首先,Y经过一维卷积进行初步特征提取得到特征e:

$$e = \mathrm{ReLU}\Big(\mathrm{BN}\Big(W_1 \cdot Y\Big)\Big), \tag{2}$$

其中,W_1指的是一维卷积变换.随后特征e会进入两个特征提取分支,左边的TDNN层通过控制空洞系数来提取动态粒度的特征e_t,右边的音素级别池化plp通过重叠滑动时间窗口提取细粒度的特征e_p:

$$e_t = \tau(e), \tag{3}$$

$$e_p = \mathrm{plp}(e), \tag{4}$$

其中$\tau(\cdot)$表示TDNN层的特征提取过程,plp(·)表示音素级池化.具体地,PLP层的输入形状为$[B, F, T]$,分别表示批次维度、频域维度和时间维度,使用一维最大池化进行池化操作,池化窗口是w,本文设为8,步长stride设为$\lceil w/2 \rceil$,池化后的输出维度为$[B, F, T']$,其中$T' = \lceil T/\mathrm{stride} \rceil$.然后,对池化后的输出进行扩维,扩展出一个新的维度,从$[B, F, T']$到$[B, F, T', 1]$,再沿这个维度复制w次,使其变为$[B, F, T', w]$,最后将这个四维张量重塑为$[B, F, T]$,并且只保留原始长度的部分,PLP层最终的输出维度为$[B, F, T]$,与输入特征保持一致.两种方式提取了不同粒度信息的特征,然后通过SE模块将这两种特征进行融合,得到e_c:

$$s = \sigma\Big(W_2 \delta\Big(W_1 \cdot g\Big[e_t, e_p\Big] + b_1\Big) + b_2\Big), \tag{5}$$

表1　MGFF-TDNN网络的详细架构表

Tab. 1　Detailed architecture of the MGFF-TDNN network

Stage	Structure	Output_size
input	-	$80 \times T$
unsqueeze	-	$1 \times 80 \times T$
conv_1	3×3,32	$32 \times 80 \times T$
ResNet-1	$\begin{bmatrix} 1\times1, 32\times6 \\ \text{dw}3\times3, 192, s=2 \\ 1\times1, 32 \end{bmatrix}$	$32 \times 40 \times T$
ResNet-2	$\begin{bmatrix} 1\times1, 32\times6 \\ \text{dw}3\times3, 192, s=2 \\ 1\times1, 32 \end{bmatrix}$	$32 \times 20 \times T$
ResNet-3	$\begin{bmatrix} 1\times1, 32\times6 \\ \text{dw}3\times3, 192, s=2 \\ 1\times1, 32 \end{bmatrix}$	$32 \times 10 \times T$
reshape	-	$320 \times T$
MTDNN Block-1	$\begin{bmatrix} \text{conv1d}\times1, 64 \\ \text{conv1d}\times3, 64, d=1 \\ \text{conv1d}\times1, 128 \end{bmatrix}\times3$	$128 \times T$
MTDNN Block-2	$\begin{bmatrix} \text{conv1d}\times1, 128 \\ \text{conv1d}\times3, 128, d=2 \\ \text{conv1d}\times1, 256 \end{bmatrix}\times6$	$256 \times T$
MTDNN Block-3	$\begin{bmatrix} \text{conv1d}\times1, 256 \\ \text{conv1d}\times3, 256, d=2 \\ \text{conv1d}\times1, 512 \end{bmatrix}\times4$	$512 \times T$
	Statistics pooling	
	FC + BN	

注:表中列出的 MTDNN 块仅列举了主要变化单位.

$$e_c = \left[e_t, e_p\right] \cdot s, \tag{6}$$

这里s是注意力权重,$\sigma(\cdot)$表示Sigmoid函数,$\delta(\cdot)$表示线性整流激活函数,$[\cdot]$是特征拼接过程,g表示全局池化操作,以获取全局表征.W_1和W_2表示一维卷积操作,b_1和b_2为偏置项.通过特征维度的缩放变换得到关于不同粒度特征对结果的贡献大小,即不同粒度特征的注意力权重.最后,通过注意力权重s对拼接特征$\left[e_t, e_p\right]$进行点乘得到融合后的特征e_c,经过一维卷积W_1进行特征提取后,与输入Y进行残差连接,最终得到特征e_o:

$$e_o = \text{ReLU}\left(Y + \text{ReLU}\left(\text{BN}\left(W_1 \cdot e_c\right)\right)\right), \tag{7}$$

e_o作为MTDNN层的输出,融合了不同粒度的特征.详细的配置如表1所示,每个M-TDNN模块由不同数量的MTDNN层组成.3个MTDNN模块分别包括3,6,4数量的M-TDNN层,每个模块的输出通道分别为128,256,512.表1中的M-TDNN模块说明了层内的主要变化单元,为了控制模型整体的参数量和计算量,本文先使用上下文为1的一维卷积层去减少特征维度,随后又在多粒度特征融合操作之后恢复到原始的特征维度,整个单元使用了跳跃连接.

3　实验及结果分析

3.1　数据集和评估指标

本文在开源说话人验证数据集VoxCeleb[22-23]上进行了实验.对于VoxCeleb,本文使用VoxCeleb2[23]的开发集进行训练,包括5994个说话者,总共1092009个语音片段,使用VoxCeleb1[22]的开发集和测试集用于评估.该数据集包括3组具有不同难度级别的评估组:VoxCeleb-O、VoxCeleb-H和VoxCeleb-E.鉴于声学环境的复杂性,本文使用噪声数据集MUSAN[24]和RIR[25]对训练数据进行了增强.

说话人验证两种常见的指标被用于模型的评估:等错误率(Equal Error Rate,EER)和目标概率为0.01的最小检测成本函数(Minimum Detection Cost Function,minDCF).等错误率是指错误接受率(False Acceptance Rate,FAR)和错误拒绝率(False Rejection Rate,FRR)相等时的识别性能.通过阈值θ来衡量两段语音是否属于同一个说话人,遍历不同的θ值,与

两个声纹特征之间的余弦相似度进行比较,计算每个阈值下的FAR和FRR,EER就是使FAR(θ)=FRR(θ)成立的阈值对应的错误率,EER越低表示说话人识别系统的性能越好.在某些应用场景中,系统对不同错误率的接受程度不一样,即对不同错误率的检测代价敏感度不一样,因此需要使用minDCF作为另一个评估指标,针对具体的场景,对不同的错误率设置不同的代价权重.最小检测成本函数综合了检测成本、先验概率和验证性能,用于评估系统在实际应用中的表现,并可以通过调整不同的参数来适应不同的应用场景.

3.2 实现细节

本实验在Intel(R) Xeon(R) Silver 4108 CPU、32GB内存和NVIDIA RTX A6000的环境下进行,使用3D-Speaker工具包[26]来训练所提出的MGFF-TDNN模型.本文使用窗口长度为25 ms、偏移量为10 ms的80维对数梅尔滤波器组(FBank)特征作为输入声学特征.除了使用噪声数据集进行增强外,还对音频应用了速度扰动,以0.9、1.0和1.1的速率随机采样,使说话人的数量增加了3倍.在实验中,采用随机梯度下降(Stochastic Gradient Descent, SGD)优化器,其初始学习率为0.1,动量为0.9,权重衰减设置为0.0001.同时集成了余弦退火调度器和线性预热调度器,用于学习率调度,最低学习率设置为0.0001.所有实验均采用加性角边缘Softmax(Additive Angular Margin Softmax, AAM-Softmax)损失[27],边缘和比例因子分别设置为0.2和32.SE块中瓶颈层的维度设置为128.最后的全连接层输出192维的说话人特征.为了提高训练效率,本文从每个音频中随机裁剪3 s的片段来构建训练样本.

在评估阶段,使用余弦相似度计算得分,该值在后端不应用分数归一化.为了证明所提出模型的有效性,本文选择的基线模型的参数和浮点计算复杂度尽可能接近或处于所提出模型的一定范围内.本文复现了基线模型ECAPA-TDNN[5]和Res2Net[8],如文献[5]所述,ECAPA-TDNN模型使用512的通道结构.

3.3 结果和分析

本文提出的MGFF-TDNN与各种基线模型进行比较的实验评估结果见表2.从表中可以看出,本文提出的方法显著优于基线模型,其中,与Res2Net模型相比,本文的方法在所有指标上都表现出显著的改进.具体来说,在3个测试集中,它的EER分别实现了35.0%、17.3%和15.5%的相对改进,minDCF的相对改进分别为38.1%、15.6%和9.7%,这是由于Res2Net侧重于局部特征建模,但在捕获不同粒度特征方面存在不足,而本文提出的方法通过引入多粒度特征建模机制,能够有效整合局部信息与全局信息,弥补这一短板.相比ECAPA-TDNN,尽管ECAPA-TDNN在语音表征中引入了通道和上下文增强机制,但其模型规模较大,参数量更高.与ECAPA-TDNN相比,本文的方法在多个测试集中以更少的参数实现了较好的性能,特别是在VoxCeleb1-O上,EER和minDCF分别提高了13.6%和23.2%.

表2 在VoxCeleb1-O、VoxCeleb1-E和VoxCeleb1-H测试集上不同网络架构的EER和minDCF性能比较

Tab. 2 EER and minDCF performancecomparison of different network architectures on the VoxCeleb1-O, VoxCeleb1-E, and VoxCeleb1-H test sets

Architecture	# Params/M	VoxCeleb1-O		VoxCeleb1-E		VoxCeleb1-O	
		EER/%	minDCF	EER/%	minDCF	EER/%	minDCF
TDNN[4]	4.62	2.31	0.322	2.37	0.273	4.25	0.393
Res2Net[8]	4.03	1.37	0.139	1.27	0.141	2.26	0.216
ResNet18[5]	13.8	1.47	0.177	1.60	0.179	2.88	0.267
ECAPA-TDNN[5]	6.19	1.03	0.112	1.06	0.122	2.05	0.201
MFA-TDNN(Lite)[18]	5.93	0.97	0.091	1.14	0.121	2.17	0.199
DS-TDNN[28]	6.50	0.90	0.118	1.15	0.140	2.11	0.199
D-TDNN[29]	2.85	1.55	0.166	1.63	0.175	2.86	0.257
D-TDNN-L[30]	6.40	1.19	0.118	1.21	0.129	2.22	0.205
MGFF-TDNN	4.78	0.89	0.086	1.05	0.119	1.91	0.195
-w/o DSM	4.81	1.14	0.117	1.17	0.128	2.11	0.206
-w/o PLP	3.52	1.04	0.114	1.08	0.126	2.00	0.196
-w/o TDNN	2.40	1.22	0.144	1.38	0.150	2.47	0.240

在消融研究中,DSM模块被移除,仅TDNN用于简单的尺寸映射,以确保与原始模型的输入维度保持一致.可以观察到,DSM在VoxCeleb1-O数据集上将EER和minDCF分别提高了21.9%和26.5%.实验

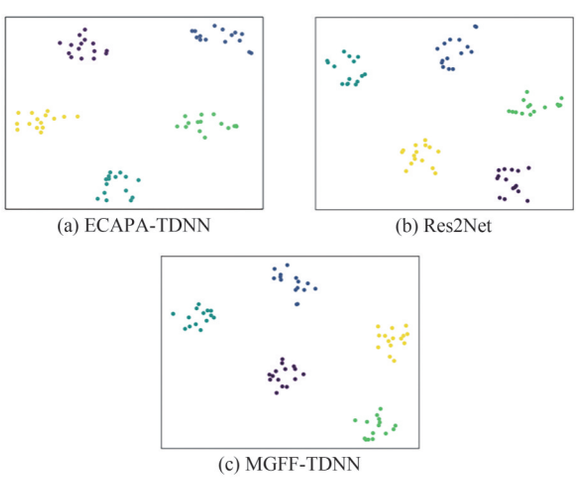
结果证明了DSM在提取时频域特征方面的有效性.去除PLP模块和TDNN模块均会导致模型在测试集中性能的下降，这强调了多粒度特征融合的重要性.

为了展示所提出模型的多粒度特征融合能力，本文在VoxCeleb测试集上评估了不同模型在3 s和5 s时长组中展示的性能，如表3所示，本文提出的MGFF-TDNN模型在大多数指标上都取得了卓越的性能.具体来说，在3 s时长组中，与Res2Net相比，MGFF-TDNN在3个测试集的EER值相对降低了24.0%、14.9%和12.0%，与ECAPA-TDNN相比也表现出一定的优势.此外，本文使用t分布随机邻域嵌入（t-Distributed Stochastic Neighbor Embedding，t-SNE）[31]可视化不同模型5 s语音时长的说话人嵌入，并比较它们的解耦能力.如图3所示，与ECAPA-TDNN和Res2Net相比，MGFF-TDNN提取的说话人嵌入在短时长语音中表现出更强的聚类能力，MGFF-TDNN使说话人嵌入更具判别性.

表3　不同模型在VoxCeleb测试集上3 s和5 s时长组的性能比较

Tab. 3　Performance comparison of different models on the VoxCeleb test set for both 3-second and 5-second duration groups

Architecture	VoxCeleb1-O				VoxCeleb1-E				VoxCeleb1-H			
	EER/%		minDCF		EER/%		minDCF		EER/%		minDCF	
	5 s	3 s	5 s	3 s	5 s	3 s	5 s	3 s	5 s	3 s	5 s	3 s
Res2Net	1.53	2.54	0.160	0.278	1.43	2.42	0.159	0.266	2.52	4.15	0.245	0.376
ECAPA-TDNN	1.16	2.05	0.143	0.279	1.21	2.10	0.139	0.235	2.30	3.87	0.220	0.358
MGFF-TDNN	1.04	1.93	0.129	0.215	1.19	2.06	0.133	0.236	2.18	3.65	0.221	0.357



(a) ECAPA-TDNN　(b) Res2Net　(c) MGFF-TDNN

图3　不同模型提取的5个说话人特征的t-SNE可视化显示

Fig. 3　t-SNE visualization of five speaker features extracted by different models

本文提出的多粒度特征融合模型不仅通过模块堆叠在垂直方向上融合了不同尺度大小的上下文信息，而且使用不同的组件在水平方向上实现了多个粒度的特征融合，这避免了单粒度提取可能会导致模型在网络加深时忽略了细粒度特征的问题.实验结果表明：提取说话人嵌入需要关注时频域特征和层内不同粒度特征的提取，有效的特征融合方法在提取具有鲁棒性且判别性强的说话人特征中发挥着重要作用.

3.4　复杂度分析

模型的复杂度可以从多个角度进行分析，包括模型性能、模型参数量和浮点计算量（Floating-Point Operations，FLOPs）等等.本节从上述角度讨论Res2Net、ECAPA-TDNN和MGFF-TDNN模型的复杂度，如表4所示.当将MGFF-TDNN与Res2Net进行比较时，尽管MGFF-TDNN的参数略多，但其FLOPs显著降低，且等错误率处于较低值；与ECAPA-TDNN相比，MGFF-TDNN在参数量和FLOPs方面都有所减少，它在VoxCeleb1-O测试集上的性能都优于这两个模型.MGFF-TDNN的计算量优势是由于使用了深度可分离卷积，极大减少了矩阵计算量，同时，使用多粒度特征融合模块强化了模型对声纹特征的建模能力，进一步提高了模型的性能.因此，在边缘设备或计算能力有限的场景中，MGFF-TDNN模型将具有更大的优势.

表4　不同模型的VoxCeleb1-O测试集上的参数量、浮点计算（FLOP）和等错误率（EER）比较

Tab. 4　Comparison of parameter counts，FLOPs，and EER on the VoxCeleb1-O test set across different models

Network	Params/M	FLOPs/G	EER/%
Res2Net	4.03	2.56	1.37
ECAPA-TDNN	6.40	1.55	1.03
MGFF-TDNN	4.78	1.49	0.89

4　结语

本文介绍了多说话人声纹识别的处理流程，基于其核心步骤，提出了一种新颖的多粒度特征融合模型MGFF-TDNN用于提取说话人特征.首先，使用二维深度分离卷积来预提取时频域特征，然后，采用多粒度特征融合模块来提取不同粒度的特征，旨

在获得更具鲁棒性的说话人嵌入,并增强可视化模型提取特征的判别性.在VoxCeleb数据集上的实验结果表明,MGFF-TDNN以更少的参数和计算复杂性实现了较好的性能.

参 考 文 献

[1] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence:IEEE, 2014: 4052-4056.

[2] RICHARDSON F, REYNOLDS D, DEHAK N. Deep neural network approaches to speaker and language recognition[J]. IEEE Signal Processing Letters, 2015, 22(10): 1671-1675.

[3] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification [C]//Interspeech 2017. Stockholm: ISCA, 2017: 999-1003.

[4] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust DNN embeddings for speaker recognition[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: ACM, 2018: 5329-5333.

[5] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]// Interspeech 2020. Shanghai:ISCA, 2020: 3830-3834.

[6] YAO J, LIANG C, PENG Z, et al. Branch-ECAPA-TDNN: A parallel branch architecture to capture local and global features for speaker verification[C]//Interspeech 2023. Dublin:ISCA, 2023: 1943-1947.

[7] ZEINALI H, WANG S, SILNOVA A, et al. BUT system description to VoxCeleb speaker recognition challenge 2019[J]. arXiv: 2019,1910.12592.

[8] ZHOU T, ZHAO Y, WU J. ResNeXt and Res2Net structures for speaker verification[C]//2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen: IEEE, 2021: 301-307.

[9] CHEN Y, ZHENG S, WANG H, et al. An enhanced Res2Net with local and global feature fusion for speaker verification[C]//Interspeech 2023. Dublin: ISCA, 2023: 2228-2232.

[10] GAO S H, CHENG M M, ZHAO K, et al. Res2Net: A new multi-scale backbone architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(2): 652-662.

[11] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141.

[12] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas:IEEE, 2016: 770-778.

[13] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted residuals and linear bottlenecks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE, 2018: 4510-4520.

[14] TAN C H, CHEN Q, WANG W, et al. Ponet: Pooling network for efficient token mixing in long sequences[C]// OpenReview. Proceedings of the International Conference on Learning Representations (ICLR). Virtual Event: OpenReview, 2022:1-8.

[15] TAMM B, VANDENBERGHE R, VAN HAMME H. Analysis of XLS-R for speech quality assessment[C]// 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz:IEEE, 2023: 1-5.

[16] 刘悦茹.基于深度学习的文本无关说话人识别方法研究[D].秦皇岛:燕山大学, 2023.

[17] SNYDER D, GARCIA-ROMERO D, SELL G, et al. Speaker recognition for multi-speaker conversations using X-vectors [C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton:IEEE, 2019: 5796-5800.

[18] LIU T, DAS R K, LEE K A, et al. MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances [C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 7517-7521.

[19] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. Lille:ACM, 2015: 448-456.

[20] JUNG Y, KYE S M, CHOI Y, et al. Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances[C]// Interspeech 2020. Shanghai:ISCA, 2020: 1501-1505.

[21] GU B, GUO W, ZHANG J. Memory storable network based feature aggregation for speaker representation learning[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2023, 31: 643-655.

[22] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: A

large-scale speaker identification dataset [C]//Interspeech 2017. Stockholm:ISCA, 2017: 2616-2620.

[23] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep speaker recognition[C]//Interspeech 2018. Hyderabad: ISCA, 2018: 1086-1090.

[24] SNYDER D, CHEN G, POVEY D. Musan: A music, speech, and noise corpus[J]. arXiv: 2015, 1510.08484.

[25] KO T, PEDDINTI V, POVEY D, et al. A study on data augmentation of reverberant speech for robust speech recognition[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans:IEEE, 2017: 5220-5224.

[26] CHEN Y, ZHENG S, WANG H, et al. 3D-speaker-toolkit: An open-source toolkit for multimodal speaker verification and diarization [C]//ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad: IEEE, 2025: 1-5.

[27] DENG J, GUO J, XUE N, et al. ArcFace: Additive angular margin loss for deep face recognition[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach:IEEE, 2019: 4685-4694.

[28] LI Y, GAN J, LIN X, et al. DS-TDNN: Dual-stream time-delay neural network with global-aware filter for speaker verification [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2024, 32: 2814-2827.

[29] YU Y Q, LI W J. Densely connected time delay neural network for speaker verification[C]//Interspeech 2020. Shanghai:ISCA, 2020: 921-925.

[30] WANG H, ZHENG S, CHEN Y, et al. CAM++: A fast and efficient network for speaker verification using context-aware masking[C]//Interspeech 2023. Dublin: ISCA, 2023: 5301-5305.

[31] VAN DER M L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605.

(责编　曹东,校对　雷建云)